KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Data Analysis of Formula 1: Top Speed Progression, Team Podium Statistics, Regression of Winning Variables

by: Zack Lee

Faculty Advisers: Professor Shelby Taylor and Professor J. Hathaway

INTRODUCTION

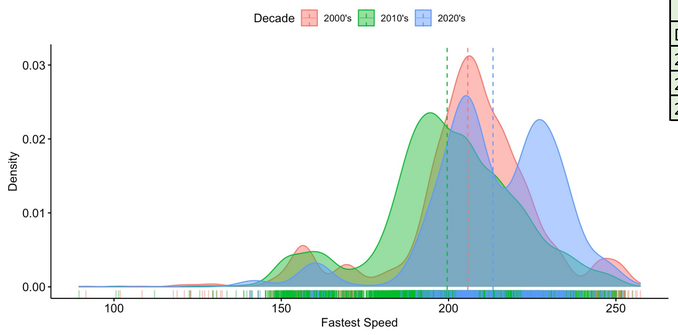
Formula 1 data from 1950 to the latest 2021 season. Formula 1(F1) is the highest class of single-seater auto racing sanctioned by the Fédération Internationale de l'Automobile (FIA) and owned by the Formula One Group. The dataset consist of all information on the Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops, and championships.

As an avid Formula 1 fan, there is a satisfaction in analyzing data over the past years of racing. I wanted do research on the progress of each team and the performance of their cars. After looking at the final dataset, I decided to analyze top speed progression, team podium statistics, and the regression of winning variables. The datasets used for this project were by searching Formula 1 statistics on Kaggle.com. Final dataset contains 25,140 rows and 25 columns of data.

METHODS

- Load Libraries (tidyverse, ggthemes)
- Load in all 13 datasets
- I joined 7 datasets together into 1 "Final_data" using SQL
    - I selected variables that would be interesting to use for analysis
- Mutate a new column called podium from information in the final position column
    - Podium is when a driver finishes in 1st , 2nd or 3rd place
- Mutate a new column called Decade from information in the year column
- Density chart of top speeds
    - Density chart of top speeds from the past few decades. This statistics only dates back to the year 2000, so we can only compare 20 years of data
    - Compare the speeds and see if the top speeds have changed over the years over the years
    - Create a table to display descriptive statistics of what is visualized in density chart
- Line Graph of Podium Counts of Teams (2015 – 2020)
    - First, I need to make a separate dataset by grouping teams and years. Then take the count of podiums
    - Are the number of podiums the same for each team?
    - I had to choose 2015 to 2020 because there were teams in Formula 1 in the past that no longer exist, and the graph would have too many teams to interpret
- Display Residual Plot to Check Normality
    - First check for multicollinearity in variables using VIF function
    - Plot residuals using lm(positionOrder ~ grid + team), final position is response variable
- Regression Table and Confidence Interval Visualization
    - Performing ANOVA to find significance in grid and team with the response variable as final position
    - Find estimates to create a regression linear formula for predicting final position based on starting postion (grid) and which team we want to predict

Graph 1: Density Chart of Fastest Speed (past 3 Decades)



Graph 2: Number of Team Podiums from 2015 to 2020

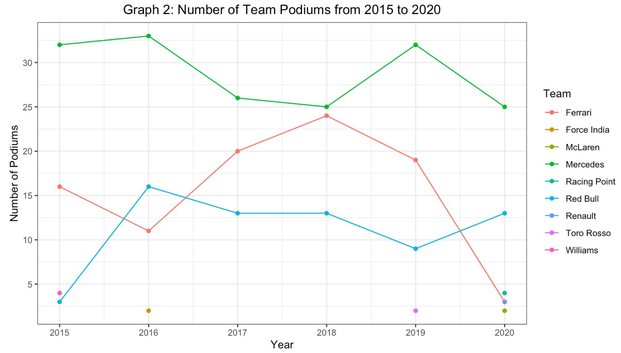
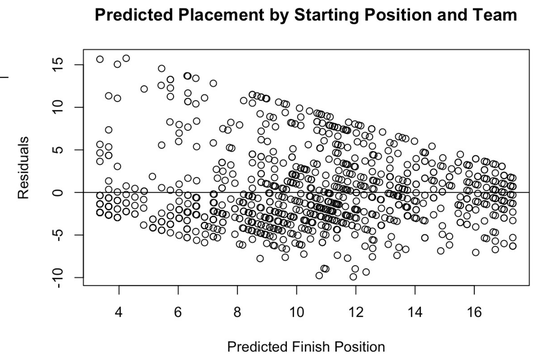


Table 1: Statistics in Top Speed in Past 3 decades

Decade	Mean	Standard Deviation	1st Quantile	3rd Quantile
2000's	206	21	198	217
2010's	200	21.1	189	213
2020's	213	19.4	204	227

Predicted Placement by Starting Position and Team



Graph 3: Confidence Intervals for Regression Analysis

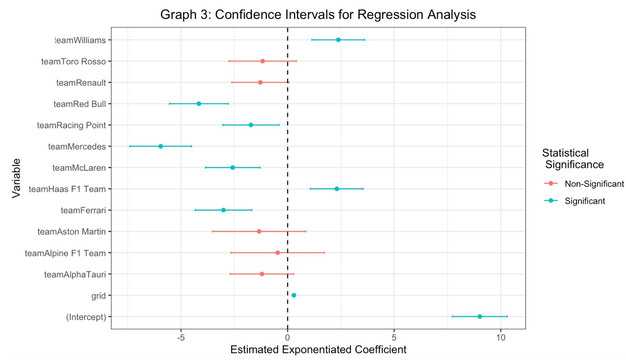


Table 2: F1 Regression Table

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	9.01	0.654	13.8	0	7.73	10.3
grid	0.295	0.034	8.81	0	0.23	0.361
team: AlphaTauri	-1.2	0.756	-1.59	0.111	-2.69	0.279
team: Alpine F1 Team	-0.466	1.11	-0.419	0.676	-2.65	1.72
team: Aston Martin	-1.34	1.11	-1.2	0.23	-3.52	0.847
team: Ferrari	-3.01	0.677	-4.44	0	-4.34	-1.68
team: Haas F1 Team	2.31	0.629	3.67	0	1.07	3.54
team: McLaren	-2.58	0.653	-3.95	0	-3.86	-1.3
team: Mercedes	-5.95	0.738	-8.07	0	-7.4	-4.5
team: Racing Point	-1.72	0.673	-2.56	0.011	-3.04	-0.404
team: Red Bull	-4.16	0.701	-5.94	0	-5.54	-2.79
team: Renault	-1.28	0.682	-1.87	0.061	-2.62	0.06
team: Toro Rosso	-1.17	0.803	-1.46	0.145	-2.75	0.405
team: Williams	2.38	0.632	3.77	0	1.14	3.62

RESULTS

Density Chart: We can see that there is somewhat of a change in top speed the past few decades, yet there isn't enough evidence to conclude that their differences are significant. The mean, 1st, and 3rd quantiles are too similar to assume a difference. **Table 1** is a table with those values. Unfortunately, the data given for this test only spans back to the year 2000, so we can not see the difference between today's top speeds compared to when Formula 1 first started in 1950.

GGPLOT (LineGraph): We can see the distribution of the number of podiums from each year since 2015. Each team does not have the same number of podiums since 2015. Clearly, Mercedes has been the most dominant F1 team since 2015. We can also see that only three teams seem to be consistently good over this time period (Mercedes, Red Bull, and Ferrari). More research can be done to find ways of why these teams are the only ones performing over the past years.

Residuals vs fitted graph: We can see that there is some skewness in residuals, yet this was slightly predicted and makes since. Let's say a driver for Mercedes starts the race in 1st place, they would be expected to finish in 4th place or better. There are instances where that the driver can get in a wreck and finish last causing a +16 residual. We can see a barrier of reality in F1 racing. A driver expected to finish in 4th can't have a -4 residual as well as a driver who is expected to finish 16th can't have a residual of +5.

Formula 1 Regression Table: We can see that grid(starting position) is significant to predicting a driver's final position. We can also see that multiple teams have been consistent enough in their performance that we can make a strong assumption of what position they will finish in. In this case, the teams with the lowest estimates are expected to finish in better positions than teams with higher estimates.

Confidence Intervals for Regression Analysis: Graph 3 is a visualization of the F1 Regression Table. With final position as the response variable, we can see that grid and multiple teams have a significance on the confidence of what place their car will finish.

R CODE

```
library(readr)
library(tidyverse)
install.packages("ggthemes")
library(ggthemes)
#Selecting Important Variables for Joining Datasets
drivernames <- drivers %>%
  select(driverId, surname)
constructors <- constructors %>%
  mutate(team = name)
constructornames <- constructors %>%
  select(constructorId, team)
#Joining Circuits to Races
racenames <- races %>%
  select(raceId, circuitId, name, year)
circuits_and_races <- circuits %>%
  dplyr::left_join(racenames,by="circuitId")
circuits_and_races1 <- circuits_and_races %>%
  select(circuitId, raceId, country, name.y, year)
#Joining Datasets to Final Dataset to use for Project
#Drivers
driver_to_result <- results %>%
  dplyr::left_join(drivernames,by="driverId")
#Constructors
constructor_to_result <- driver_to_result %>%
  dplyr::left_join(constructornames,by="constructorId")
#Status
status_to_result <- constructor_to_result %>%
  dplyr::left_join(status,by="statusId")
#Final dataset
Final_data <- status_to_result %>%
  dplyr::left_join(circuits_and_races1,by="raceId")
#Mutating Decades for final data
Final_data <- Final_data %>%
  mutate(Decade = if_else(year >= 1950,
                          paste0(year  %/% 10 * 10, "'s"),
                          paste0((year - 1900) %/% 10 * 10, "'s")))
#Adding Podium to final dataset
Final_data <- Final_data%>%
  dplyr::mutate(podium = if_else(positionOrder == 1 | positionOrder == 2 |positionOrder == 3, "Podium", "No Podium"))
Final_data

#Writing into Excel
write.csv(Final_data,"F1_Final_data.csv", row.names = FALSE)

#Now we can Begin Data Analysis
best_drivers <- Final_data %>%
  group_by(team,year) %>%
  count(podium)
best_driver1 <- best_drivers %>%
  filter(podium == "Podium")%>%
  filter(n > 1)%>%
  filter(year >= 2015, year < 2021)
#Graph 2
ggplot(best_driver1, aes(x = year, y = n))+
  geom_point(aes(x = year, y = n, colour = team))+
  geom_line(aes(x = year, y = n, colour = team))+
  theme_bw()+
  labs(x= "Year", y = "Number of Podiums", title = "Graph 2: Number of Team Podiums from 2015 to 2020")+
  theme(plot.title = element_text(hjust=0.5))+
  guides(color = guide_legend("Team", order = 1))+
  scale_y_continuous(breaks = c(5,10,15,20,25,30,35))

#Chart 1
Final_data %>%
  dplyr::group_by(Decade) %>%
  filter(year > "1999")%>%
  dplyr::summarize(Fastest_mean = mean(fastestLapSpeed,na.rm=T),Fastest_sd = sd(fastestLapSpeed,na.rm=T),
                   Fastest_1quantile = quantile(fastestLapSpeed, probs = 0.25, na.rm = T),
                   Fastest_3quantile = quantile(fastestLapSpeed, probs = 0.75, na.rm = T))

#Change lapspeed to numerical value
Final_data$fastestLapSpeed <- as.double(Final_data$fastestLapSpeed)
#Graph 1
Final_data%>%
  filter(year > "1999")%>%
  ggpubr::ggdensity(x="fastestLapSpeed", add="mean", rug= TRUE,
            color="Decade", fill="Decade")+
  labs(title = "Graph 1: Density Chart of Fastest Speed (past 3 Decades)", x = "Fastest Speed", y = "Density")+
  theme(plot.title = element_text(hjust=0.5))

Final_data$grid <- as.numeric(Final_data$grid)
Final_data_reg <- Final_data %>%
  filter(year > "2018")
F1_mod<- lm(positionOrder~grid + team , data=Final_data_reg)
#Check Multicollinearity
car::vif(F1_mod)
# Regrssion Table
F1_mod %>%
  modernditve::get_regression_table()
#Residual Plot
F1_residuals = resid(F1_mod)
plot(fitted(F1_mod), F1_residuals,
     ylab = "Residuals", xlab = "Predicted Finish Position",
     main = "Predicted Placement by Starting Position and Team")
abline(0,0)

#Graph 3
F1_mod %>%
  broom::tidy(conf.int=T,exponentiate=T) %>%
  dplyr::mutate(Sig = if_else(p.value < 0.05,"Significant",
                              "Non-Significant")) %>%
  ggplot(aes(x=term,y=estimate)) +
  geom_errorbar(aes(ymin=conf.low,ymax=conf.high,color=factor(Sig)),width=0.1) +
  geom_point(aes(color=factor(Sig))) +
  geom_hline(aes(yintercept = 0),color="black",linetype="dashed") +
  labs(x = "Variable",
       y = "Estimated Exponentiated Coefficient",
       color = "Statistical \n Significance") +
  theme_bw() + coord_flip() +
  ggtitle("Graph 3: Confidence Intervals for Regression Analysis") +
  theme(plot.title = element_text(hjust=0.5))
```